Adversarial Robustness of ChatGPT Using Adversarial Attacks

Pratik Karmakar
Department of Computer Science
National University of Singapore
pratik.karmakar@u.nus.edu

Jason Xinyang Zhang
Department of Computer Science
National University of Singapore
e0983283@u.nus.edu

Parul Bansal
Department of Computer Science
National University of Singapore
parul.bansal@u.nus.edu

Félix Chavelli
Department of Computer Science
National University of Singapore
chavelli@comp.nus.edu.sg

Abstract

The robustness of large language models has been a topic of interest for researchers and practitioners. In this study, we investigate the robustness of ChatGPT 3.5 and Google's FLAN T5[3] by subjecting them to Seq2Sick attacks and the Adversarial General Language Understanding Evaluation (Adversarial-GLUE)[9] dataset. We measure the effectiveness of the attacks by comparing the Bleu scores, Meteor scores, and Euclidean distances between the original and perturbed outputs. Our results show that both models are susceptible to Seq2Sick attacks and that the Adversarial-GLUE dataset is effective in exposing weaknesses in their performance. The perturbed outputs from Flan T5 model had lower Bleu and Meteor scores and higher Euclidean distances wrt to the original outputs than the perturbed outputs from ChatGPT 3.5. Thus, the robustness of ChatGPT was found to be higher than that of Flan T5. These findings highlight the importance of evaluating the robustness of large language models and the need for developing better defense mechanisms against adversarial attacks. Our study treats both the models as Black-Boxes (which these are indeed) and using a black-box attack method, provides insights into the strengths and weaknesses of ChatGPT 3.5 and Flan T5, which can be used to improve their performance and robustness in the future. Overall, our study contributes to the growing body of research on the robustness of large language models and provides valuable insights for researchers, practitioners, and developers working on natural language processing applications.

1 Introduction

With the widespread use of natural language processing (NLP) systems, there is a growing concern about the vulnerability of these systems to adversarial attacks. Adversarial attacks refer to the deliberate manipulation of input data to trick a machine learning model into making incorrect predictions or outputs. Such attacks could have serious consequences in various applications, including chatbots and conversational agents.

ChatGPT 3.5 and Google Flan T5[3] are two of the most popular pre-trained language models that are widely used in various NLP applications. However, their adversarial robustness against attacks remains unexplored. In particular, it is unclear how these models compare to each other in terms of adversarial robustness, especially against the Seq2Sick attack and the AdvGLUE dataset.

The scope of this paper is to compare the adversarial robustness of ChatGPT 3.5 and Google Flan T5 against the Seq2Sick attack and the AdvGLUE dataset. The Seq2Sick attack is a type of adversarial attack that aims to generate malicious input sequences that can 'fool' the large language models. On the other hand, the AdvGLUE dataset consists of various adversarial examples that can evaluate the robustness of language models against various types of adversarial attacks.

To compare the adversarial robustness of ChatGPT 3.5 and Google's Flan T5, we will perform the following tasks:

- Generate adversarial examples using the Seq2Sick attack on both models and evaluate their performance on these examples.
- Test both models on the AdvGLUE (qqp) dataset and compare their performances on adversarial examples.
- Analyze the results and draw conclusions on the adversarial robustness of ChatGPT and Google Flan T5 against the Seq2Sick attack and the AdvGLUE dataset.

Overall, the paper aims to provide insights into the comparative adversarial robustness of ChatGPT and Google Flan T5, which can help researchers and practitioners in selecting the most robust pre-trained language models for their NLP applications.

2 Related work

Our proposed evaluation framework for the adversarial robustness of ChatGPT and FLAN T5 builds upon a growing body of research on the vulnerability of deep neural networks to adversarial attacks, and the efforts to develop more robust models. Wang et al.[9] propose a new benchmark dataset, AdvGLUE, for evaluating the generalization performance of natural language understanding models. The dataset includes a range of adversarial examples designed to test the robustness of language models, providing a valuable resource for evaluating the effectiveness of our proposed framework. Moradi et al.[6] present a robustness comparison of BERT, RoBERTa, XLNet an ELMo on 5 different tasks: Text Classification, Sentiment Analysis, named Entity Recognition, Semantic Similarity check and Question Answering. The authors use their crafted character and word level perturbations for the robustness test. In a recent and remarkable work Wang et al.[10] compare the adversarial and out-of-distribution robustness of ChatGPT against other existing models mostly on classification tasks and one translation task. Bang et al.[1] present a multi-tasking (8 tasks) efficiency aspect of ChatGPT using different existing datasets. Guo et al.[4] leverage the Human ChatGPT Comparison Corpus (HC3) dataset to evaluate the closeness of ChatGPT and human performances and understanding of language. Jeblick et al.[5] address the question of reliability of ChatGPT in medical field, specifically in radiology and their results show that even if in most cases ChatGPT produces factually right report, the factually incorrect reports are potentially harmful.

3 Background

3.1 ChatGPT

ChatGPT is a large language model developed by OpenAI, based on the GPT-3.5 architecture. It is designed to generate coherent and contextually relevant responses to user input. ChatGPT is one of the most advanced conversational AI models in existence, with 175 billion parameters that enable it to generate highly sophisticated and nuanced responses.

3.2 FLAN T5

FLAN T5[3] is a large-scale language model developed by researchers at Google. It is based on the T5 architecture, which was originally introduced by Google Research in 2020 as a unified framework for natural language processing (NLP) tasks. FLAN T5 builds on this architecture by incorporating additional training data and fine-tuning techniques to achieve state-of-the-art performance on a range of NLP benchmarks.

3.3 Robustness

Model robustness refers to the ability of a machine learning model to perform well under different conditions, such as changes in the distribution of the input data, the presence of noisy or adversarial inputs. There are several mathematical techniques that can be used to quantify the robustness of a model. For classification models the measures of robustness can be accuracy, precision, recall, F1 score, ROC curves, and adversarial attacks. For regression models, one can use the R^2 value or MSE loss.

In this work, the two models under study are generative in nature. Thus we use two methods to quantify the robustness:

- Text level similarity
- Vector level similarity

Let x be the original input x' be the perturbed input to the model, and Sim be a similarity metric. Then we claim that model f_1 is more robust than model f_2, if

$$Sim(f_1(x), f_1(x')) > Sim(f_2(x), f_2(x'))$$

3.4 Seq2sick

Seq2Sick[2] is a type of adversarial attack that targets sequence-to-sequence (seq2seq) models, which are commonly used in natural language processing (NLP) tasks such as machine translation, text summarization, and dialogue generation. Text inputs are represented in discrete domain and word level perturbation to texts to keep the meanings close is a tricky task. To craft adversaial samples, seq2sick model aims to solve an optimization problem as below:

$$\min_{\delta} L(X + \delta) + \lambda \cdot R(\delta)$$

Where $R(\cdot)$ is the regularization on the perturbation δ and $L(\cdot)$ is the loss to penalize unsuccessful attacks. While using this, it is common to find the adversarial example vectors located in a region with no embedding vectors. To solve this problem the authors use a projection method and the objective finally becomes:

$$\min_{\delta} L(X + \delta) + \lambda_1 \cdot R(\delta) + \lambda_2 \cdot \sum_{i=1}^{N} \min_{w_j \in W} \{||x_i + \delta_i - w_j||\}$$

From seq2sick, we use the *non-overlapping* attack. It requires that the words of the output of the perturbed input should not match the words of the original input.

To use seq2sick attack we use the Textattack[7] library's implementation of the attack which uses a grid search method instead.

3.5 AdvGLUE

Adversarial GLUE Benchmark (AdvGLUE) is a comprehensive robustness evaluation benchmark that focuses on the adversarial robustness evaluation of language models. It covers five natural language understanding tasks (Sentiment Analysis, Duplicate Question Detection (QQP) and Natural Language Inference) from the famous GLUE tasks and is an adversarial version of GLUE benchmark. AdvGLUE considers textual adversarial attacks from different perspectives and hierarchies, including word-level transformations, sentence-level manipulations, and human-written adversarial examples, which provide comprehensive coverage of various adversarial linguistic phenomena.

We use a small subset of the QQP corpus for our experiment which has 363,846 examples in the train set and 442 examples in the test set. The adversarial perturbation consists of Word level, sentence level and human crafted examples. The word and sentence level perturbations have a combination of Typo-based, Embedding-similarity-based, Context-aware and Knowledge-guided perturbations. The average word length is 4.234 characters and average sentence length is 7.623 words.

4 Methodology

Black-Box: In case of both models, they are treated as black-box models. We do not use access to the parameters of the models, instead their APIs are used. Thus we have access to only the inputs and the outputs of the models.

In this work to compare the robustness of the two models, we use two different methods:

- Seq2sick
- AdvGLUE

We use questions from WebQuestions dataset and questions from the AdvGLUE dataset to query the models and evaluate their robustness.

Seq2sick: We use seq2sick attack as mentioned in 3.4 to perform a black-box attack on the model to measure the robustness. In this method, the attacker sends numerous queries to the model to find the adversarial examples to fool the models under attack. Here fooling refers to a classifier misclassifying usually, but as we are working on generative models, a successful 'fooling' refers to perturbing the input to generate entirely different output by the model. In Table 5 we present the comparison of amount of words replaced to attack the two models under study, the average number of queries and the success rate of the attack.

Model	Words replaced (%)	Avg. queries	Attack Success (%)
ChatGPT	16.45	59.87	95.00
FLAN T5	16.65	60.79	96.82

Table 1: Comparison of seq2sick attack on the two models

AdvGLUE: The QQP corpus of AdvGLUE contains pairs of sentences. Among the pairs, some are generated using adversarial attacks and some are human generated. All the pairs are human supervised to ensure that the meanings of the questions do not change. We thus use these pairs of questions to measure the robustness of the models.

The two methods differ slightly in the fact that when we use the seq2sick method, the amount of perturbation is not in our control. Thus the perturbation in case of the two models are not necessarily equal. While we use questions pair from the AdvGLUE dataset, the pairs of questions used in both models are same thus the input perturbation is same for both.

We evaluate the robustness using the variation of output in two domains as mentioned in 3.3. To compare evaluate the results in textual domain, we use BLEU1-4 and METEOR as similarity metrics. To perform vector level evaluation, we use sentence transformers to transform the sentences into embedding vectors and then find the Euclidean distance between them (which can be considered as inverse of similarity).

5 Experimental Results

We applied two kinds of attacks on two large state-of-the-art language models and compared their stability performance with metrics.

Table 2 presents the mean BLEU1-4 scores, mean METEOR scores and mean Eclidean distances between pairs of original and perturbed questions and original and perturbed answers by the models.

Figure 1 and Figure 2 show the BLEU1-4 and METEOR score between the original sentence and the perturbed sentence under AdvGLUE Attack and Seq2sick Attack. We can see that the ChatGPT output has larger BLEU1-4 and METEOR than FlAN T5 output under both AdvGLUE Attack and Seq2sick Attack. This means that the ChatGPT model has a more stable output as the perturb results match the original outputs better.

Figure 3 and Figure 4 show the mean and standard deviation of the Euclidean Distance between the original sentence and the perturbed sentence under AdvGLUE Attack and Seq2sick Attack. We can see that the ChatGPT output has a smaller mean Euclidean Distance than FlAN T5 output under

ChatGPT		BLEU1	BLEU2	BLEU3	BLEU4	METEOR	Euclidean Distance (mean)
seq2sick	Input	0.825	0.713	0.542	0.324	0.836	0.635
	Output	0.195	0.125	0.091	0.067	0.306	0.938
AdvGlue	Input	0.693	0.582	0.411	0.317	0.756	0.602
	Output	0.452	0.346	0.295	0.25	0.5	0.623
FLAN T5							
seq2sick	Input	0.828	0.713	0.536	0.346	0.838	0.631
	Output	0.04	0.012	0.005	0.003	0.046	1.045
AdvGlue	Input	0.693	0.582	0.411	0.317	0.756	0.602
	Output	0.455	0.217	0.159	0.105	0.371	0.684

Table 2: Comparison of robustness of ChatGPT 3.5 and FLAN T5

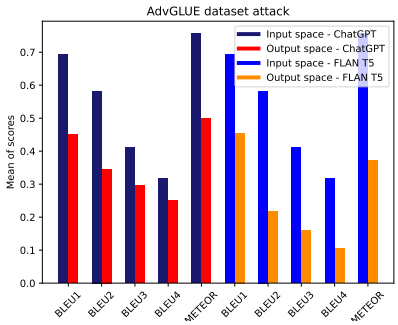

Figure 1: AdvGLUE Attack Metrics Score

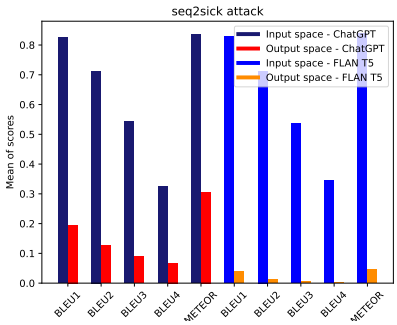

Figure 2: Seq2sick Attack Metrics Score

both AdvGLUE Attack and Seq2sick Attack. This means that the ChatGPT model has a more stable output in the embedding aspect.

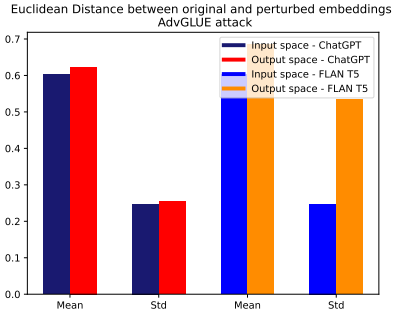

Figure 3: AdvGLUE Attack Euclidean Distance

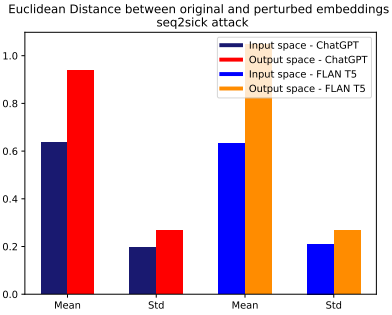

Figure 4: Seq2sick Attack Euclidean Distance

6 Discussion

It appears that for the two adversarial attacks performed (either seq2sick or advGLUE), ChatGPT is less affected than FLAN T5. Indeed, METEOR and BLEU1 to BLEU4 scores are always less disturbed by the attacks in the case of ChatGPT than in the case of FLAN T5. Beyond the textual domain, this observation is corroborated by the analysis of the distances between the original and perturbed embeddings which are in both cases smaller in the case of ChatGPT than in the case of FLAN T5. This leads to the conclusion that ChatGPT seems to be more robust and secure than FLAN T5.

During our experiments we identify two different problems with the methods that we use or have been used in the literature.

- When we use the seq2sick method, even after using the formulation mentioned in the original research, we see that it sometimes generates perturbed inputs which are not close by the

meanings. Thus using this method as an automated attack against these models may not always be fruitful for robustness analysis.

- When we use the question pairs from the AdvGLUE dataset, while they are meaningfully close, the perturbations are not model specific. Thus the attack would generally have less rate of success.

7 Future Work

Future work in comparing the robustness of large language models against adversarial attacks include several avenues. First, it is possible to investigate how new models perform in improving LLM robustness. It has recently been shown that adversarial training with MixUp augmentation can improve LLM robustness against adversarial attacks [8]. Another idea could be to assess the ability of models to generalize to out-of-distribution examples or to detect adversarial examples. Finally, new adversarial attacks beyond Seq2Sick and AdvGLUE could be developed to further evaluate the robustness of LLMs. For instance, attacks that target specific semantic properties of text or attacks that aim to manipulate the output of a model in a more subtle way could be developed.

References

[1] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[2] Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3601–3608, 2020.

[3] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[4] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

[5] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*, 2022.

[6] Milad Moradi and Matthias Samwald. Evaluating the robustness of neural language models to input perturbations. *arXiv preprint arXiv:2108.12237*, 2021.

[7] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, 2020.

[8] Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. Better robustness by more coverage: Adversarial training with mixup augmentation for robust fine-tuning. *CoRR*, abs/2012.15699, 2020.

[9] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021.

[10] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023.